Australian Bureau of Statistics

2961.0 - Census Working Paper 93/2 - Self-coding, 1996
Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 06/01/1994

Summary

Main Features

Census Working Paper 93/2

COMPARISON OF SELF-CODED AND WRITE-IN RESPONSES: JULY 1992 TEST

CONTENTS

LIST OF TABLES

INTRODUCTION

This working paper will assess the effect of using a Census form based on a self-coding rather than write-in format by comparing coded responses from the two forms used in the July 1992 Census test. The Optical Mark Recognition (OMR) form required responses to be self-coded while the Optical Character Recognition (OCR) form required responses to be written in. This is the first time that extensive analysis of the difference between the pattern of self-coded and write-in responses has been possible using parallel forms. Previously, data for the 1986 Census, which was mostly write-in, was compared with test data, which was mostly self-coded. The July 1992 Test was conducted with a view to assessing the possibility of using OCR technology in the 1996 Census but it has since been decided that OMR technology will be used as in 1991. The data will still be useful, however, in comparing written and self-coded responses. It should be noted that there were several problems affecting the quality of the test data. These are documented in 'July 1992 OCR Test, Report on Form Design Issues'.

Background

Previous work had shown that there was a possible bias in self-coded responses, known as a 'list effect'. Prior to the 1986 Census, the possibility of using a self-coding format for an ethnic origin/ancestry question was assessed. It was concluded that 'the provision of response options influences both respondents' perceptions of a question's meaning and the answers provided' and thus introduced a bias. Some of the causes of these influences noted were:

- people chose a category from the list of response options in preferences to one not on the list
- the response options encouraged responses different from those which would have been provided without them
- people would not read the entire list - they would tick the first option they came to that seemed suitable.

(**The Measurement of Ethnicity in the Australian Census of Population and Housing,** Cat. No. 2172).

The only clear indication of a possible 'list effect' found for the 1991 Testing Program was in the Religion question (see Census Working Paper 91/5). In particular, self-coding affected responses of 'No religion'. It was thought that the change in form design was most likely to affect respondents without strong religious affiliation, possibly because the list of religious denominations may have brought additional social pressure on their responses. Due to the limitations of the Testing Program, analysis of list effects for other variables was not conclusive.

July 1992 Census Test

The July 1992 Census Test was conducted on 21 July 1992 in Brisbane. The areas included were: New Farm, Spring Hill, West End, South Brisbane and Highgate Hill. OMR and OCR forms were distributed to alternate dwellings. Details of the number and type of dwellings involved, and of the total number of people included on forms received, are given in Table 1 overleaf.

Table 1: Number of Dwellings and Persons included in the July 1992 Census Test

	OMR form	OCR form

Total Dwellings	2,492	2,482
Occupied Dwellings	1,728	1,726
Non-contact	143	135
Refusal	255	280
Mailback	61	56
Unoccupied	305	285
Total Persons	3,042	3,140

1991 Census data showed that the July 1992 Test area had a large proportion of recently arrived migrants: the proportion of people who were born in overseas countries was very high in the Test area (36.7% of persons in the Test area were born outside Australia while only 19.3% of persons in Queensland were) as was the proportion whose State of Usual Residence One Year Ago was overseas (4.1% in the Test area compared to 1.8% in Queensland overall). The July 1992 Test OMR data gave similar percentages for the area, which indicates that there continued to be a high proportion of recently arrived migrants.

Method and Results

This report focuses on variables where the write-in versus self-coded contrast occurs. The responses from both forms in the July 1992 Test have been tested for significant differences, both for the overall distribution (using Standard Chi-Square analysis and Multinomial Logistic Regression) and between each category (using T-tests). The response distributions for each of the questions examined and the results of statistical testing are included in Attachment 1, along with copies of the questions as they appeared on the July 1992 Test OMR and OCR forms.

The differences between the overall distributions were significant for all variables discussed except Birthplace and Birthplace of Father. The overall distribution for Age was also tested but was not significant. Since there did not appear to be any list effect for Age it will not be considered further in this report. The data for Age is included in Attachment 2.

The discussion is presented in five sections with some of the questions grouped for analysis because of similarities in the type of response categories offered on the OMR form. The first of these sections includes questions related to ethnic origin: Birthplace, Birthplace of Father, Birthplace of Mother and Language. The second includes questions for which the OMR form had a list of ordered categories: Income, Year of Qualification, Year of Arrival and Hours Worked. The remaining three sections all examine only one question as each has a unique format or response pattern. These questions are: Age Left School, State of Usual Residence One Year Ago and Religion.

DISCUSSION OF RESULTS

Birthplace, Birthplace of Father, Birthplace of Mother, Language

For each of these questions, the OMR form offered a list of seven choices and then an 'Other' category with space for a write-in answer. The OCR form offered a self-coded response for the most common answer (Australia/English) and a write-in space for other answers.

The differences between the responses given on the July 1992 Test OMR and OCR forms followed a similar pattern for all four variables, although the differences were only significant for Birthplace of Mother and Language. Table 2 below summarises the distribution of responses for Birthplace of Mother (the complete list of responses can be found in Attachment 1). The distribution for Language was similar.

Table 2: Summary of responses for Birthplace of Mother

	OMR form		OCR form	
Response	No.	%	No.	%
Australia	1,518	52.3	1,540	53.2
Other categories listed on OMR form	646	22.2	578	20.0
Other	741	25.5	778	26.9
Total	**2,905**	**100.0**	**2,896**	**100.0**

The comparison of OCR and OMR responses indicates that there may have been a 'list effect' as people tended to mark a listed category rather than write in a response. This appears in two ways. Firstly, in the higher proportion of people who marked the single listed category on the OCR form than the same category on the OMR form and, secondly, in the higher proportion of people who marked the listed categories on the OMR form than wrote in the equivalent responses on the OCR form.

The proportion of people who marked the listed response category 'Australia' was around 1% greater (which is significant) for Birthplace of Mother on the OCR than the OMR form. Such a difference tended to occur, to a lesser extent, for the other three questions and supports the idea that some people responded differently to the OCR form than they would have to the OMR form, choosing to mark the single listed category rather than write in a response.

The proportion of people who wrote a response at the 'Other' category was smaller for the OMR Test form, compared to the proportion writing corresponding responses on the OCR form, for all of the above variables except Birthplace. This difference was not statistically significant for any variable, but can be seen particularly clearly in the Language question (where there may be a complicating factor). Table 3 summarises Language responses.

Table 3: Summary of responses for Language

	OMR form		OCR form	
Response	No.	%	No.	%
English	2,099	72.4	2,153	73.6
Cantonese	85	2.9	27	0.9

Mandarin	39	1.4	17	0.6
Other categories listed on OMR form	299	10.3	297	10.1
Other	376	13.0	432	14.8
Total	**2,898**	**100.0**	**2,926**	**100.0**

The proportion of people responding 'Cantonese' and 'Mandarin' was significantly higher on the OMR form (on which these categories were listed) than on the OCR form. It would appear that the presence of 'Cantonese' and 'Mandarin' in the list on the OMR form may have helped people to understand the detail of response required. It is possible that 'Other' responses on the OCR form include responses of 'Chinese' which could be more common in the absence of prompts of 'Cantonese' and 'Mandarin'.

The apparent list effect (where people were more likely to mark a listed response) was also observed in testing prior to the 1991 Census. The proportion of people giving responses in the 'Other' categories on OMR forms tended to be lower in comparison to the (non-OMR) 1986 Census for Birthplace, Birthplace of Father and Birthplace of Mother. The greater than expected proportion of people marking a self-coded response may have been due to people being prompted by the list when determining a response or selecting a self-coded response which appears suitable rather than writing in a non-listed response.

Age Left School

The difference between the distribution of responses for the July 1992 Test OMR and OCR forms was statistically significant for this variable. The categories which had the greatest differences were 'Did not go to school' and '19 years or older'.

On the July 1992 OMR Test form, 1.6% of people said they did not go to school, while 6.5% of people on the OCR form gave this response. This could be due to the design of the OCR form, which gave the options 'Still at primary or secondary school' and 'Did not go to school' and then had space for Age Left School to be written in. The question could have appeared to offer only the first two choices so people may have responded 'Did not go to school' meaning they did not attend school in 1992.

The difference observed for '19 years or older' supports other analysis of 1991 Age Left School data included in the report '1991 Census: Age Left School Data Quality Overview'. A large increase in the response '19 years or older' has been observed in all States between the 1986 and 1991 Censuses. While some difference would have been expected as the result of rising retention rates, the increase appears to be at least partly due to form design.

In the July 1992 Test, the proportion of people stating '19 years or older' on the OMR form (8.8%) was much higher than on the OCR form (3.0%). This may be due to people misinterpreting the term 'school' and including tertiary institutions or misunderstanding the question and giving their current age rather than their age when they left school. The report mentioned above explains these possible effects in more detail.

State of Usual Residence One Year Ago

There was a significant difference between the distribution of responses on the July 1992 Test OMR and OCR forms. Table 4 below summarises the differences between the two forms. The difference in the proportion of people not stating a response was tested separately from the coded responses.

Table 4: Summary of responses for State of Usual Residence One Year Ago

Response	OMR form		OCR form		Difference
	Persons		Persons		
	No.	%	No.	%	%
NSW / VIC	106	3.5	77	2.5	1.0
QLD	2,557	84.8	2,452	79.8	5.0
Other States/Territories	19	0.6	21	0.7	-0.1
Elsewhere	116	3.8	202	6.6	-2.8
Not Stated	217	7.2	322	10.5	-3.3
Total	**3,015**	**100.0**	**3,074**	**100.0**	

On the OMR form, there was a significantly larger proportion of people who stated their State of Usual Residence One Year Ago as Queensland and a smaller proportion of people who stated 'Elsewhere' or didn't respond. It appears that there may be several form design issues here.

Some people who had not moved interstate in the last year may not have bothered writing in a response on the OCR form, perhaps seeing it as irrelevant, whereas people were more likely to respond to the self-coded question. This would have led to lower non-response and a larger proportion responding 'Queensland' on the OMR form.

The OCR form gave the option of writing in the three letter abbreviation of the State or Territory or marking 'Elsewhere'. As observed earlier, people appear to have preferred marking a self-coded response rather than writing in an answer and this could have led to the greater proportion of people marking 'Elsewhere' on the OCR form than the OMR form. Also, people may have marked 'Elsewhere' or not responded on the OCR form if they did not know, or were not sure of, the abbreviation required for the States and Territories.

The format for the responses on the OMR form, with the States and Territories listed first and 'Elsewhere' included at the bottom, may have led people to mark the most appropriate State or Territory, even if this was incorrect. They may not have realised that 'Elsewhere' was a possible response. This could have led to a larger proportion of people marking States or Territories on the OMR form and fewer marking 'Elsewhere'. This effect would have been more noticeable in the Test area than in other areas as it is known that there was a large population of migrants, and so there were more people for whom the 'Elsewhere' option was appropriate. Confusion could also have arisen from the different meanings assigned to 'Elsewhere' in Q8 (Usual Residence) and Q9 (Usual Residence One Year Ago). While this occurred on both forms, it would have been more obvious on the OCR form and could have led to people marking 'Elsewhere' to indicate that they moved in the last year (rather than they lived overseas one year ago as was meant).

Religion

The question on Religion tends to be very sensitive to changes in form design as it is optional and the information required is both personal and subjective. It was noted in the testing program for the 1991 Census that it was affected more than other questions by the change from a write-in response to a self-coded response. In the 1986 Census, an instruction was included to direct people who had no religion to write 'None'. In the August 1989 Test, which used self-coded rather than write-in responses, no such instruction was included and the proportion of people marking the 'No Religion' category which was at the end of the list decreased markedly. An instruction was included in the 1991 Census and this appears to have resolved this problem.

Large differences in the overall distributions were also observed in the July 1992 Test between the OMR and OCR forms. These differences are summarised in Table 5 below. Again, the difference in the proportion of people not stating a response was tested separately to the coded responses.

Table 5: Summary of responses for Religion

Response	OMR form		OCR form	
	Persons		Persons	
	No.	%	No.	%
Listed denomination	1,733	57.2	1,244	41.7
Other	323	10.3	741	24.9
No religion	591	19.5	315	10.6
Not stated	384	12.7	681	22.8
Total	**3,031**	**100.0**	**2,981**	**100.0**

A significant difference noted between the 1986 Census results and the August 1989 Test was the decrease in the proportion of people giving responses coded to 'Other'. This was also observed in the July 1992 Test. The proportion of people who marked 'Other' was much smaller on the OMR form than the OCR form, while the proportion in each of the listed categories (except Greek Orthodox) was much greater. This may indicate that, on OMR forms, people either marked a response that seemed appropriate or were prompted by the list to give a response they would not have given otherwise.

The proportion of people who stated that they had 'No religion' on the OMR form was much greater than that on the OCR form. Both forms included a specific instruction directed at those without a religious affiliation. Separate testing of non-response indicated that the level of non-response was significantly different between the OMR and OCR forms. This indicates that people without a religious affiliation who would not have bothered to answer a write-in question may have marked the 'No religion' response on the OMR form. Thus, the presence of this response in the list may have resulted in a lower non-response rate. Census results also show a lower non-response rate for Religion in the 1991 Census compared to the 1986 Census.

Another factor which may have contributed to the lower level of non-response on the OMR

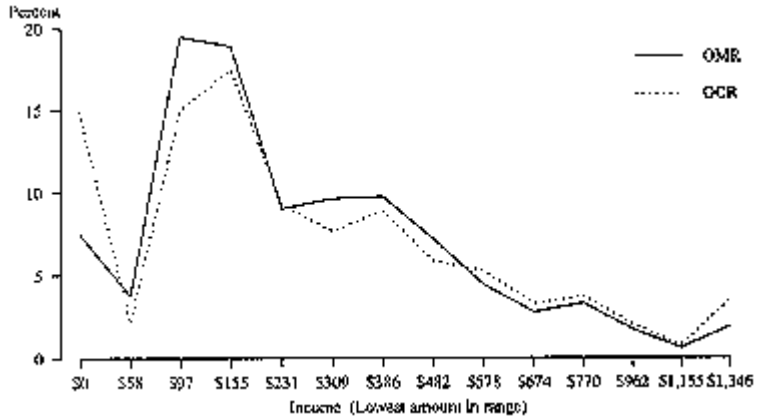
form, compared to the OCR form, may be that some people with a weak religious affiliation did not bother to write-in a response on the OCR form.

Income, Year of Qualification, Year of Arrival, Hours Worked

The response categories for these four questions were completely self-coded on the OMR form and appeared as an ascending numerical scale. There did appear to be a list effect as, for each of the variables, the greatest differences between the July 1992 Test OMR and OCR forms were for the top and bottom categories. The differences for each question are discussed below.

The graph below shows the distribution of responses for Income for the OMR and OCR forms. This graph shows that the difference in the proportion of responses for the OMR and OCR forms was greatest for the lowest categories (Less than $58, $58-$96 and $97-$154), for $309-$385, and for the highest category (More than $1346). It also shows that, although the difference was not always significant, for categories $578-$673 and higher, the proportion of people was consistently greater on OCR forms than OMR forms for these categories.

DISTRIBUTION OF RESPONSES FOR INCOME



This indicates that there was a list effect, possibly in three ways. Firstly, it is possible that people were less likely to mark the extreme categories in the OMR list (that is, the highest or the lowest). Secondly, the presence of the list could have given people a concept of 'high' and 'low' incomes and may have influenced some people (for example those concerned about imagined comparisons with income declared to the Taxation Office) to mark a lower response than they would have without the list. This second effect was not supported by significant differences but the proportion of responses on the OCR form in these categories tended to be higher. Thirdly, on the OCR form people were offered one self-coded category of 'Nil Income' and so, as observed for other questions, people may have preferred to mark this category rather than write in a response. Another form design issue involved concerns the list of inclusions. On the OMR form they are listed on separate lines and are easy to read, while on the OCR form they are joined together in a single paragraph and much more difficult to read. This may have resulted in more accurate answers on the OMR form, although there is little direct evidence in the data to support this.

Despite the very different natures of the questions on Year of Qualification and Year of Arrival, the pattern of differential response between the OMR and OCR forms was similar. The format of the response required for both questions was a year. Table 6 below summarises the distribution of Year of Qualification. The difference in the proportion of people not stating a response was tested separately to the coded responses.

Table 6: Summary of responses for Year of Qualification

Response	OMR form		OCR form	
	Persons		Persons	
	No.	%	No.	%
Before 1971	233	23.4	281	27.8
1971-1989	515	51.7	490	48.4
1990-1992	181	18.2	132	13.0
Not stated	67	6.7	109	10.8
Total	**996**	**100.0**	**1,012**	**100.0**

For **both** questions the only significant differences were for the lowest (Before 1971) and highest (1990 to 1992) categories. The proportion of people responding 'Before 1971' was higher on the OCR form while the proportion responding '1991 to 1992' was higher on the OMR form. The non-response rates for these questions were also higher on the OCR than the OMR form. This may indicate that people may have been more likely to guess from the categories offered on the OMR form or they may have been helped by the categories to understand the question or formulate an answer where they otherwise would not have answered. That similar patterns were found in the responses to very different questions supports the idea that the differences were due to a list effect rather than differences between the sample populations.

For the Hours Worked question, the proportion of people responding for the two extreme categories ('None' and '49 hours or more') was significantly higher on the July 1992 Test OCR form than the OMR form. The proportion of people responding '35-39 hours' and '41-48 hours' was significantly higher on the July 1992 Test OMR form than the OCR form. Although the difference was not significant, the proportion of people who stated '40 hours' was greater on the OCR form, possibly because people rounded off the number of hours they worked.

The fact that the proportion of people classified to the extreme levels of these four variables is significantly lower for the OMR form than the OCR form for all but '1990-1992' of Year of Qualification and Year of Arrival indicates that, for variables where a choice is made from an ordered list, people may be more reluctant to choose extreme categories than those in the middle.

CONCLUSIONS

The comparison of the OMR and OCR forms used in the July 1992 Test revealed several questions for which responses may be subject to a 'list effect' as a result of the self-coding format of the OMR form. The effect appears to vary according to the nature of the question. The main points emerging from the analysis are:

- The responses on the OMR form may be biased by people tending to mark a given category rather than write a response under 'Other', possibly because it was easier to mark a given category, or because they were prompted by the listed categories. This effect was observed for the questions on Birthplace, Language and Religion. A similar tendency was sometimes found on the OCR form as some people appeared to prefer

to mark the one listed category rather than write in a response.

- The comparison of responses to Age Left School from OMR and OCR forms also supported analysis of 1991 Census results that indicated that the increase in the proportion of people responding '19 years or more' to the OMR question was contributed to by the use of a self-coding rather than a write-in form.
- Responses on the OMR form may also be affected by people tending to mark responses at the top of the list rather than the bottom. This may have led to a smaller proportion of people on OMR forms responding 'Elsewhere' to the State of Usual Residence One Year Ago question as this category was at the bottom of the list.
- Another factor which may have affected responses for Income, Year of Qualification, Year of Arrival and Hours Worked could be a tendency for people not to mark the most extreme categories on an ordered list.

About this Release

ABOUT THIS RELEASE

This working paper will assess the effect of using a Census form based on a self-coding rather than write-in format by comparing coded responses from the two forms used in the July 1992 Census test. The Optical Mark Recognition (OMR) form required responses to be self-coded while the Optical Character Recognition (OCR) form required responses to be written in. This is the first time that extensive analysis of the difference between the pattern of self-coded and write-in responses has been possible using parallel forms. Previously, data for the 1986 Census, which was mostly write-in, was compared with test data, which was mostly self-coded. The July 1992 Test was conducted with a view to assessing the possibility of using OCR technology in the 1996 Census but it has since been decided that OMR technology will be used as in 1991. The data will still be useful, however, in comparing written and self-coded responses. It should be noted that there were several problems affecting the quality of the test data. These are documented in 'July 1992 OCR Test, Report on Form Design Issues'.
